INTRODUZIONE A INTERNET
per studenti universitari
3-2: ricerca di informazione in WWW

Dr. Giorgio F. Signorini

Dipartimento di Chimica

Università di Firenze

signo@chim.unifi.it

http://www.chim.unifi.it/~signo

4 febbraio 2010

Copyright 2

INDICI DEL WWW **3**

Materiale che si trova su WWW 4
Indici globali 5
Indici per soggetto e per parola 6
Indici o cataloghi per soggetto 7
Indici per parola 8
Ricerca avanzata 9
Motori di ricerca 10
Lista di motori di ricerca 11
Motori di ricerca “intelligenti” 12
Rilevanza 13
Importanza intrinseca 14
Il PageRank di Google 15
Altri algoritmi e metodi per la misura dell’importanza 16
Altri motori di ricerca 17
Google 18
Google directory 19
Indici specializzati 20
Wiki 21
Wikipedia 22

RICERCHE **23**

Strategie di ricerca 24
Obiettivi 25
Strumenti 26

ATTENDIBILITÀ **27**

Come valutare l’attendibilità dell’informazione 28
Un possibile schema di valutazione della qualità dell’informazione 29
Informazioni contenute nell’URI 30

Autorevolezza 31
Indicatori di qualità 32
indicatori di qualità 2 33
indicatori di qualità 3 34

BIBLIOGRAFIA **35**
Bibliografia: ricerca informazione su WWW 36

INDICI DEL WWW

Materiale che si trova su WWW

- WWW può essere visto “come” una grande biblioteca, analizzabile con strumenti di ricerca automatici
  - non è organizzato dall’alto, né omogeneo
  - insieme a fonti autorevoli, molto materiale di scarso valore o attendibilità; come distinguere?
  - contiene materiale spontaneo, non filtrato (è facile pubblicare il vostro romanzo su WWW)
  - inglese! [a]

[a] Secondo una statistica (fonte non più reperibile), nel 2000 il 68% del WWW era scritto in inglese; solo l’1.6% in Italiano. [http://global-reach.biz/globstats/refs.php3]

Indici globali

- Non esiste un indice “ufficiale” del WWW
- Nessun indice può essere completo
  - ci sono ca. 72 milioni di server WWW al mondo (9/2005, cfr. `http://news.netcraft.com/archives/2005/09/index.html`), alcuni con migliaia di pagine
  - ogni giorno moltissimi documenti nascono, muoiono, cambiano, o cambiano indirizzo (URI)

5 / 36

Indici per soggetto e per parola

- Esistono due tipi di “indice” del WWW:
  1. per soggetto
  2. per parola
- Raccolta di documenti (URI) da inserire nell’indice:
  - scelta fatta dai compilatori (es. Encyclopaedia Britannica, Open Directory)
  - su notifica da parte degli autori (es. LookSmart)
  - a pagamento (es. Overture)
  - automatica (es. AltaVista, AllTheWeb.com, Google)

6 / 36

Indici o cataloghi per soggetto

- C'è sempre una qualche scelta di chi compila il catalogo:
  - schema di classificazione è arbitrario
  - assegnazione a (una o più) classi è arbitraria (può usare indicazioni fornite dall'autore)
- Queste scelte possono essere affidate anche ad un software specializzato
- Classificazione è generalmente gerarchica.
  Es: `computer > software > games`

7 / 36

Indici per parola

- Indici per parola sarebbero troppo grandi da consultare direttamente.
- La pagina di un indice contiene una casella dove inserire le parole da cercare. Avviando la ricerca, un programma sul sito remoto ricerca le parole nell'indice.
- Come risultato, si ottiene un elenco di pagine che contengono le parole cercate.

8 / 36

Ricerca avanzata

- Ricerca “avanzata”:
  - operatori logici e di posizione (AND, OR, NOT, NEAR)
  - frasi intere
  - frammenti di parola
  - campi particolari (titolo, URI,...)
  - proprietà (data, lingua,...)
  - cercare in un sito particolare

9 / 36

Motori di ricerca

- Motore di ricerca *(“search engine”):* un programma che visita periodicamente (quasi) tutto il WWW e registra il contenuto delle pagine (più altre informazioni) in un indice.
- Normalmente si intende: motore di ricerca = la sua sede (home-page) su WWW, che permette l’accesso all’indice
  - per parola (sempre)
  - per soggetto (se c’è)
- Se una pagina viene aggiornata, entro un certo tempo il motore di ricerca aggiorna corrispondentemente l’indice

10 / 36

Lista di motori di ricerca

- Pricipali motori di ricerca:

google!	`http://www.google.com`
Yahoo!	`http://www.yahoo.com`
Ask.com	`http://search.ask.com`
Windows Live	`http://www.live.com`

11 / 36

Motori di ricerca "intelligenti"

- Un motore di ricerca primitivo elenca tutte le risorse che contengono le parole cercate, in un ordine casuale
  - In questo modo spesso le risposte utili sono mescolate a quelle inutili
  - si può migliorare restringendo la ricerca, per esempio aggiungendo alcune parole chiave
  - però in questo modo si possono perdere risposte utili
- Un sistema migliore è ordinare le risposte per valore, in modo che le più utili siano elencate tra le prime. Un motore di ricerca "intelligente" misura il valore di una risposta in due modi:
  1. la **rilevanza** della risorsa rispetto alla ricerca fatta
  2. l'**importanza** intrinseca della risorsa
- Altre forme di intelligenza: correzione a supposti errori di battitura (Google), etc.

12 / 36

Rilevanza

Per valutare la rilevanza di una risorsa nei confronti della ricerca fatta si può tenere conto, ad esempio, di:

- numero di ripetizioni delle parole nel documento
- posizione delle parole (è meglio se compaiono nel titolo, o comunque in cima al documento)
- vicinanza tra le parole
- ordine tra le parole è quello richiesto

Altre forme di intelligenza utili per migliorare la rilevanza:

- suggerimento di ricerche simili a quella data (Lycos)
- indicazione di categorie (soggetti di un catalogo) pertinenti alla ricerca data

13 / 36

Importanza intrinseca

- Le pagine possono essere classificate in base alla loro importanza, indipendendemente dalla rilevanza rispetto ad una particolare ricerca
  - ad es. una pagina dell'Istituto Superiore di Sanità è più "importante" del blog di uno sconosciuto
- È difficile fare una valutazione obiettiva (e automatica) dell'importanza delle pagine. Ci sono alcuni criteri:
  - quanti volte sono visitate (popolarità "pura")
  - quante altre pagine la citano (una sorta di popolarità "esperta")
  - La misura di popolarità "di ordine superiore" inventata da Google (PageRank)
- Altre misure di importanza
  - se il titolo contiene il termine "Home Page", etc.

14 / 36

Il PageRank di Google

- PageRank, la misura di importanza di una pagina inventata da Google ne ha fatto in pochi anni il motore di ricerca più usato (circa il 50%)
- È basato sulla popolarità "esperta". Una pagina A è importante:
  - se molte pagine B contengono link ad A
    - se le pagine B sono **a loro volta** importanti, cioè sono citate da molte pagine C
- Questo metodo è obiettivo, ma non è inattaccabile. Si possono fabbricare delle "bombe per Google" in modo da fare apparire tra i primi risultati un sito voluto
  - esempio famoso: la ricerca "miserable failure" con Google fornisce la biografia di G. W. Bush al primo posto. Provate "fallimento" o "buffone".

Altri algoritmi e metodi per la misura dell'importanza

- Dato che PageRank non è "a prova di bomba", Google usa anche altri metodi per la misura dell'importanza
  - In altri metodi si usano altri parametri e/o altri algoritmi
- In generale gli algoritmi sono pubblici, ma le ricette per la loro combinazione sono segrete
  - "si dice" che Google usi *centinaia* di metodi
- Altri motori di ricerca usano un metodo principale diverso da PageRank: HITS (Teoma), SALSA, etc[a]

[a] vedi ad es.: Langville A. M., Meyer C. D., *A survey of eigenvector methods for Web information retrieval*, SIAM review 2005, 47 (1) 135 -161

Altri motori di ricerca

- Gli elementi di cui è costituito un motore di ricerca sono
  - il database
  - il sistema di campionamento del WWW per costruire e aggiornare il database
  - il sistema di ricerca e di ordinamento risultati
- Ci sono molti indici, ma gli elementi fondamentali sono quasi sempre derivati da uno dei seguenti:
  - Google
  - Yahoo
  - Teoma / Ask.com
  - Windows Live Search
- Questi usano una “tecnologia proprietaria” (segreta)
- Ci sono tecnologie “open-source” (ht://Dig, Nutch, etc.), in cui tutto il funzionamento è noto. Mozdex.com è un sito sperimentale che utilizza Nutch

Google

Google! (`http://www.google.com`) è un motore di ricerca intelligente. È attualmente considerato il migliore[a],[b]

1. ricerca base:
   - tutte le parole
   - risposte in ordine di valore
   - ci sono anche risposte a pagamento
   - si può accedere alla “copia cache”
2. ricerca avanzata permette di specificare meglio
3. ha anche un catalogo per soggetto

[a] S. Dennis, P. Bruza, R. McArthur, Web searching: A process-oriented experimental study of three interactive search paradigms, JASIST (Journal of the American Society for Information Science and Technology) , 53,2,pp.120-133 (2002)

[b] Griesbaum, J., *Evaluation of three German search engines: Altavista. de, Google. de and Lycos. de,* Information research Vol. 9 No. 4, July 2004 http://informationr.net/ir/9-4/paper189.html

Google directory

- Il catalogo per soggetto (Directory) di google è http://dir.google.com
- include solo una piccola parte[a] delle pagine del WWW, recensite e classificate da volontari (Open Directory)
- le pagine di ogni categoria sono ordinate per importanza come i risultati dell'indice per parola (motore di ricerca) di google
- collegamento tra directory e motore di ricerca:
  - sulla directory si può cercare con il motore di ricerca
  - sui risultati del motore di ricerca è riportata la categoria corrispondente della directory (se c'è)

19 / 36

[a] 3.800.000(Open Directory)/1.500.000(base google) [10/2002]

Indici specializzati

- Google contiene una miriade di ricerche specializzate:
  - News (notizie)
  - Scholar (articoli su riviste scientifiche)
  - Print (libri stampati)
  - Froogle (oggetti in vendita)
  - ...
- Altri hanno solo una ricerca specializzata: Kelkoo (oggetti in vendita)
- Ricerca parallela su più database specializzati: A9 (libri, immagini, ... fino a 200)

20 / 36

Wiki

- Con un browser WWW è possibile non solo *leggere informazione creata da altri*, ma anche *collaborare con altri per crearla*
- uno *wiki* è una pagina che può essere letta, ma anche *modificata* da qualunque visitatore
- l'idea è quella di favorire la collaborazione volontaria
- si pensa che questo modo di costruire informazione sia in molti casi più efficiente del sistema tradizionale basato su compilatori esperti
  - naturalmente è possibile anche sabotare una pagina intenzionalmente: ma si pensa che per ogni atto di pirateria ci sono molti collaboratori pronti a ripristinare il contenuto "buono"
- esempio di successo: wikipedia

Wikipedia

- Wikipedia (`http://www.wikipedia.org`) è un'enciclopedia on-line costruita e aggiornata dai lettori
- Contiene articoli in tutti i campi, particolarmente nell'informatica e nella scienza e tecnica
- Esistono Wikipedie in tante lingue. Anche quella in italiano è abbastanza esauriente
- Gli articoli sono on-line e possono essere rintracciati da un motore di ricerca

Strategie di ricerca

1. Ricerca per soggetto:
   - è sconsigliata perché
     - cataloghi sono incompleti e poco aggiornati
     - difficile trovare categoria giusta
2. Ricerca per parola:
   - Scopo ideale:
     - SOLO risposte rilevanti
     - TUTTE le risposte rilevanti (non sempre)
   - Ragionevole approssimazione:
     - 10-100 risposte
     - le più rilevanti tra le prime

Obiettivi

- Possibili obiettivi di una ricerca su WWW (in ordine di precisione decrescente)
  - pagine che contengano un testo preciso [es. legge, letterario]
  - pagine che contengano un oggetto non-testuale preciso [es. programma, riproduzione di un quadro]
  - pagine che *si riferiscano* ad un oggetto preciso [es. homepage ditta, orario FS]
  - un'informazione precisa nel contenuto ma non nella forma [un dato storico o scientifico, un'informazione tecnica o commerciale]
  - una "bibliografia" generale su un argomento

Strumenti

- Possibili strumenti
  - ricerca per parola, diretta:
    - con motore di ricerca intelligente (che ordini le risposte per rilevanza e importanza)
    - per frase, per parole molto inconsuete, per host/URI
    - per parole con operatori logici e di posizione
  - ricerca per parola, indiretta:
    - ricerca di pagine di accesso a database off-line [es. dizionari], poi ricerca su quelle
    - ricerca di cataloghi specifici o “nodi di scambio”, poi uso di quelli
    - ricerca di pagine simili a quelle cercate per trovare nuove parole chiave
  - ricerca per soggetto:
    - su cataloghi settoriali
    - su cataloghi generali ma selettivi (es. Britannica)
    - facendo anche ricerca per parola all’interno di un soggetto (es. Google Directory)

26 / 36

ATTENDIBILITÀ

27 / 36

Come valutare l’attendibilità dell’informazione

- I principali motori di ricerca, come Google, danno precedenza a pagine “importanti”
- Non è però garantito che l’informazione sia attendibile al 100%. Ad esempio
  - casi in cui tra i primi risultati compaiono pagine non compilate da esperti
  - informazione che varia rapidamente
- Gli algoritmi di misura dell’importanza fanno -diciamo- il 90%. Il resto dipende dal nostro giudizio.
- L’attendibilità dell’informazione è basata su reputazione dell’autore (“Chi è che l’ha detto?”) più che sul numero di conferme (a volte, informazioni sbagliate sono ripetute molte volte)

28 / 36

Un possibile schema di valutazione della qualità dell'informazione

- Autore e fonte (qualità *a priori*)
  - Identificabilità
    - Identità dell'autore/fonte
    - fonti dell'informazione
    - forma di pubblicazione
  - Credibilità
    - credenziali
    - assenza di motivazioni esterne
- Qualità dell'informazione (*in sé*)
  - Contenuto
  - Forma
- Sostegno esterno (qualità *a posteriori*)

29 / 36

Informazioni contenute nell'URI

- Il nome a domini che appare nell'URI a volte dà indicazione sulla fonte.
- Es: una pagina di "sbufalamento" di un presunto virus (GoodTimes): `http://www.hr.doe.gov/goodtime.html`.
- Possiamo crederle, se osserviamo il nome del server:
  - I domini di primo e secondo livello (`.doe.gov`) dicono che si tratta di
    - un sito governativo USA
    - precisamente, il Ministero dell'Energia (Department Of Energy, DOE)
- Ma attenzione! ci sono siti con nomi ingannevoli:
  - `http://www.whitehouse.org` finge di essere il sito ufficiale della Casa Bianca USA
  - `http://gatt.org` finge di essere il sito ufficiale del GATT, trattato che è alla base del WTO

30 / 36

Autorevolezza

1. una fonte autorevole dovrebbe contenere:
   - nome dell'autore
   - qualifica dell'autore
   - appartenenza dell'autore a enti o società
   - data di creazione
2. ci sono anche altri indicatori di qualità

31 / 36

Indicatori di qualità

1. Credibilità
2. Accuratezza
3. Ragionevolezza
4. Sostegno esterno

32 / 36

indicatori di qualità 2

- Credibilità:
    - qualifica e/o titoli (accademici, professionali,...) dell'autore o dell'organizzazione
    - l'autore è facilmente contattabile, anche per verificare quanto sopra
    - l'organizzazione prevede un controllo di qualità sull'informazione (revisione, tipo riviste scientifiche)
    - l'informazione è pubblicata su libro (!)
- Accuratezza:
    - aggiornamento
    - concretezza (verificabilità)
    - completezza

33 / 36

indicatori di qualità 3

- Ragionevolezza:
    - tono equilibrato
    - mancanza di interessi diretti
    - assenza di specifiche richieste o pressioni sul lettore
- Sostegno esterno:
    - recensioni indipendenti
    - documento contiene riferimenti

34 / 36

Bibliografia: ricerca informazione su WWW

- Ed Krol, *The Whole Internet,* 2nd Edition
- Ed Krol, *Internet,* Milano, Jackson libri, 1994.
- D. J. Barrett¸*NetResearch: Finding Information Online,* Songline and O'Reilly & Associates¸1997¸ (ISBN 1-56592-245-X)
- F. Metitieri, R. Ridi, *Biblioteche in rete: istruzioni per l'uso,* Roma [etc.] : GLF editori Laterza, 2005 (ISBN 88-420-7663-5). Online su http://www.laterza.it/bibliotecheinrete/index.htm
- *Google Corporate Information: Technology*, http://www.google.com/corporate/tech.html
- Vaughn Aubuchon, *Google ranking factors,* http://www.vaughns-1-pagers.com/internet/google-ranking-factors.htm
- R. Harris, *Evaluating Internet Research Sources,* 17/11/1997. http://www.virtualsalt.com/evalu8it.htm
- Ballard Spahr Andrews & Ingersoll, LLP, *Evaluating the Quality of Information on the Internet,* 27/10/2005 http://www.virtualchase.com/quality/index.html